The information age is burying business in information. To the rescue come firms like Verity, with systems for targeted text retrieval.

Haystack searching

By David Churbuck

YOU CAN BLAME George Boole, the 19th-century logician, for the frustration computer users suffer trying to extract the perfect piece of information from a database of documents.

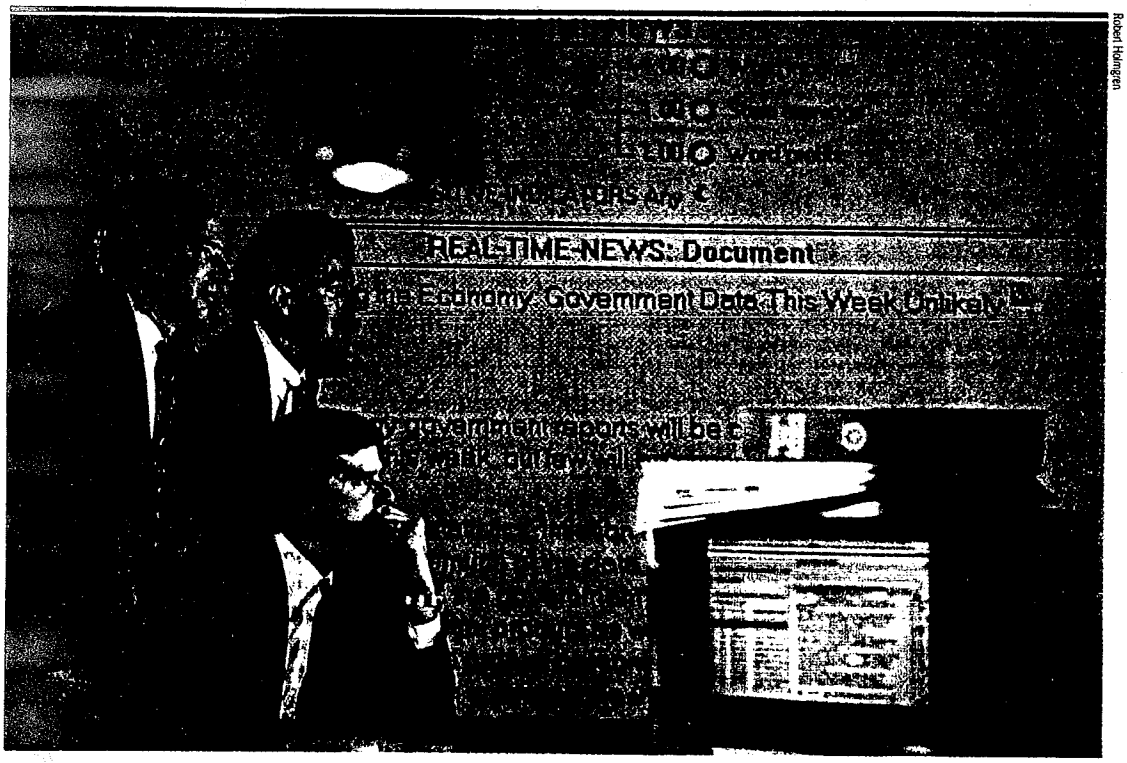
Boolean logic, which frames queries with ORs and ANDs, is fine if the document sought can be precisely targeted by key words, dates or places of publication. But often it can't be.

Let's suppose that you want to assemble a folder of newspaper and magazine articles addressing George Bush's reelection prospects. Too restrictive a search may miss important articles. If you limit the retrieval to items that mention both the President and "reelection," for example, you will miss the ones that talk about his popularity and the coming election but don't contain the word "reelection." Too broad a query also gives useless results. Ask for all items containing "George Bush OR President Bush" and one popular news retrieval service responds that there are 1,866,525 selections.

The answer to this needle-in-the-haystack problem is intelligent text retrieval, software that can combine the raw power of a computer shuffling through tens of millions of documents with the common sense that a human researcher would bring to the problem. Most text retrieval being done today is still of the literal-minded, Boolean variety. But as intelligent searching systems become more powerful, they could expand the market considerably, from one where legal, academic and scientific users predominate to one where corporations routinely search published documents to find out more about their competitors and customers. Analyst Ann Palermo of International Data Corp. projects that the market for retrieval software will quadruple by 1995, to $400 million a year.

Among the players in the smart searching business are Fulcrum, an Ottawa company that sells text retrieval software to other software and hardware companies, and Verity Inc., a Mountain View, Calif. company that has a license to commercialize some of Advanced Decision Systems Research's techniques. Verity has its roots in Advanced Decision, now a subsidiary of management consultants Booz, Allen & Hamilton, which contracts with the government.

Clifford Reid, the 33-year-old MIT



Robert Holmgren

Verity President Michael Pliner with vice presidents Clifford Reid and Robert Williams **Their software turns computers into high-speed reference librarians.**

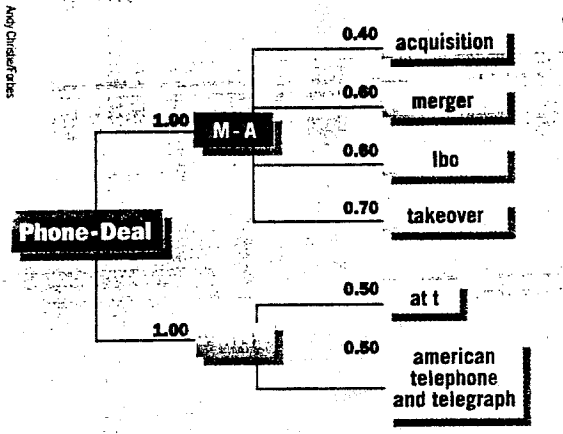
Fuzzy retrieval

Classic text searching is very rigid in its Boolean OR and AND logic. Verity's Topic software, in contrast, allows a researcher to hint around about his topic of interest and fetch documents that merely come close. The software is a descendant of the "fuzzy logic" school of artificial intelligence.

The drawback to Topic is that it demands more mental effort on the part of the user. But, having drafted a query on a particular subject, the user can apply it again and again—say, to each day's wire service stories.

In the example below, a stock analyst needs a system for retrieving items relating to corporate takeovers by AT&T. He composes a query he calls "Phone-Deal." The query branches off into two subtopics. The first lists words suggestive of corporate dealmaking; the second, the different ways AT&T could be represented in news articles.

The results of this search range from an item about Teradata's second-quarter results that describes AT&T's purchase of that company to an item mentioning AT&T and an "acquisition"—a word that turns out to refer to a government purchasing program.



Andy Christ/Forbes

Score	Headline
0.88	Teradata Reports Second-Quarter Results
0.80	Firms Dispute FCC Approval of Bell/Metro Merger
0.80	NCR Sales Channels Now Marketing the StarServer
0.80	AT&T Reports Earnings
0.75	AT&T Unit Unveils High-Speed Transmission Technology
0.60	BellSouth Reports 1991 Results
0.50	AT&T Waives Sign-Up Fee For Reach Out Ohio Plans
0.40	Pentagon Extends CHAMPUS Contract; Re-Bid Anticipated

and Harvard Business School graduate who founded Verity in 1988, calls his firm's approach "conceptual searching." Verity is aiming its Topic software not at occasional users of the sort who might call up Dialog or Dow Jones News Retrieval when they visit a library, but rather at corporations making the same sorts of inquiries over and over. Example: a pharmaceutical company tracking adverse reactions to its products through several databases connected over a local area network. One database would hold Federal Drug Administration reports on the drugs, another internal lab reports and another communications from physicians noting reactions.

Topic takes advantage of the repetitiveness of queries by picking the user's brain for information about the subjects that are relevant and their relation to one another. An expert familiar with the subject of the search assigns weights to search terms, then composes a sample query on adverse reactions that would link together terms that might be germane. Whenever a person needed to research the subject of adverse reactions, he would run the all-purpose query in conjunction with the name of a specific drug.

Verity's customers include the White House, which uses it to route wire service reports to the appropriate staffers; Bankers Trust, which uses it to study competitors; and the Defense Department agency charged with ensuring that sensitive or dangerous technology isn't shipped to the wrong hands. "We have a large volume of data that needs to be sifted through to track those people who may be shipping restricted technology to the bad guys," says this agency's Colonel Francis Wilson. "We just can't do that in a timely fashion by having analysts go through it. With Topic, we can filter the information based on parameters that set the importance of certain terms."

Reid explains: "In a Boolean system you issue a query and in essence segment the database into two sets, those documents that match and those that don't. Our system establishes another set, the set that broadly includes everything you might be interested in, with the system determining the degree to which the document is in your set."

Like other retrieval systems being proposed in academia or developed by commercial firms, Verity's can rank retrieved documents in importance. "Topic is like telling the researcher to give you everything on a subject but put the good stuff on top," says Robert Williams, vice president of marketing at Verity.

Topic can also make connections between documents that fit the user's query and loosely related documents. Says Williams: "Often the most interesting result of a search isn't finding the right document but finding 12 that are related in an interesting way, inspiring a point of view that you hadn't thought of before."

That's getting more sophisticated, but until smarter systems such as Topic become more popular, computers will still have a way of constantly reminding the user that they are just computers. When Reid used the old-fashioned technique to query a database for articles about earthquakes, he got back a lot of irrelevant baseball stories. The search program had concluded from the many mentions of the interrupted World Series game in stories about the 1988 San Francisco quake that the Oakland A's and the San Francisco Giants had some causal connection to the subject.

So maybe machines can never function as good researchers without outside help from humans. But, given the huge volume of text to be searched, neither can humans function without the machines. ■